Spring 2019

Do Men Matter? In Statistics, Probably

Michael Kelly
Western Washington University

Mike Kelly

5/25/2019

Senior Project

Do Men Matter? In Statistics, Probably

Introduction:

In statistical genetics, there are several parameters of a dataset which a researcher might want to know, but are difficult to estimate in practice. In this paper, we will be focusing on allele frequencies, null alleles, inbreeding coefficients and, to a certain extent, beta values. A common technique for obtaining these values, developed by Amy Anderson and her co-workers, is to jointly estimate all of them using an EM-algorithm and the method of maximum likelihood. Despite this technique being effective in general, it is currently unable to deal with males at X-linked markers. The purpose of this project is to modify the existing algorithm to deal with males at X-linked markers, and conduct simulation testing to see the effect of these unusual markers and identify possible uses. In this paper, we work from the ground up to give readers a basic understanding of the statistics and genetics fundamentals required to grasp the EM-algorithm and subsequent modifications. We also include a discussion of the modifications, as well as the results of our simulations.

Intro to Genetics:

Before we dive into things like null alleles and inbreeding coefficients, it is important that we have some background in the basics of genetics. At the core, we will be looking at loci (the plural of locus) which are specific points in an individual's DNA where genes are read. When we refer to markers on DNA or markers on a chromosome, we are really referring to loci.

At each locus, an individual will have a number of alleles: generally two but, as we will see later, not always. These alleles are usually represented as letters. The specific combination of alleles at a particular locus determines the traits of the individual for that locus. For example, suppose fur color in a breed of dog is determined by one specific locus with possible alleles 'A' and 'a'. Having AA at that locus might mean the dog has solid brown fur while aa means solid

white fur and Aa means brown fur with white stripes. Note that Aa and aA are interchangeable and both simply referred to as Aa.

An individual is homozygous at a locus all alleles are the same at that locus. In the example above, a dog with AA or aa would be considered homozygous at the fur color locus. On the other hand, an individual is considered heterozygous at a locus if they have different alleles at that locus. A dog with Aa would be heterozygous.

Alleles are passed down from parent to child. In general, an individual gets one allele from their father and one allele from their mother at every locus. Which allele they get is essentially random. If a dog is Aa at the fur color locus, then its' child is just as likely to get the 'A' allele as it is to get the 'a' allele.

An individual's specific combination of alleles at a certain locus is known as their genotype. AA, Aa, and aa are all possible genotypes for the dog fur example. However, each genotype does not necessarily represent a different visible trait. It is possible, for example, that both AA and Aa give a dog solid brown fur (possibly because 'A' is the dominant allele). The observed trait of an individual for a particular locus is known as the phenotype. Brown fur and white fur are both phenotypes in our example.

Sex-linked loci, markers on the X-chromosome, behave differently from the autosomes discussed above. Note that for the purposes of this project, we will only be considering mammals, and will be defining an individual as female if they have two X-chromosomes and male if they have one X-chromosome and one Y-chromosome. In any event, markers on the X-chromosome are interesting because males only have a single allele while females have two alleles. This will be very important to us later. For example, suppose whether or not our dogs have spots is controlled by a locus on the X-chromosome. Any dog with an 'A' allele has spots whereas any dog with no 'A' allele has no spots (A is dominant). Then females have possible genotypes AA (spotted), Aa (spotted), aa (not spotted) while the males have possible genotypes A (spotted) and a (not spotted).

Hardy-Weinberg Equilibrium:

Now we start dipping our toes into a bit of math. A common and very useful concept in statistical genetics is that when everything behaves "nicely", you can treat any given allele as a

random draw from all the alleles in the population. This concept is known as Hardy-Weinberg Equilibrium (HWE).

Suppose we are looking at a particular locus for some population which has two possible alleles; A and a as before, although the alleles themselves don't really matter for the concept. Suppose also that the population in question engages in entirely random mating. Then the population will be in HWE.

Let the frequency of the 'A' allele (or the probability of 'A' coming up if a random allele was selected from the population, denoted P(A)) be p_A, and let the frequency of the 'a' allele (P(a)) be p_a. Suppose now that we drew a random individual from our population (perhaps we get a guy named Fred) and looked at their genotype at the locus in question. What is the chance Fred is AA? What about Aa or aa? Since we have entirely random mating and no other funny business, we can think of Fred's alleles as being two random, independent alleles from the population. Thus, the chance Fred is AA is just the probability we get an 'A' followed by another 'A'. This would put P(AA) at $p_A^* p_A$, or p_A^2. Similarly, P(aa) = p_a^2. As for Aa, there are two possible ways Fred could get it: he could get an 'a' then an 'A', or he could get an 'A' then an 'a'. Thus, P(Aa) = $2*p_A*p_a$. Any population in which the genotype frequencies behave as if the alleles were random, independent draws from the population is said to be in HWE. This will happen if the population is randomly mating.

If you have a dataset consisting of genotypes of randomly-chosen individuals from a population in HWE, it is then very easy to estimate the (population) frequencies of each allele. However, things are rarely so nice. There are several common factors which can interfere with HWE. Some of the big players are inbreeding and population structure, each of which can cause an excess of homozygous genotypes compared to HWE. Furthermore, even if your population is in HWE, we can still have the appearance of excess homozygosity due to a certain type of problematic alleles known as null alleles. As we will see later, this project will revolve around estimating allele frequencies, null allele frequencies, and inbreeding coefficients jointly.

Inbreeding Coefficients:

As discussed previously, one of the main things which can cause Hardy-Weinberg Equilibrium to fail is the presence of inbreeding. In order to understand how inbreeding works

mathematically, we need to discuss the concept of Identical By Descent (IBD) and kinship coefficients.

At a particular locus, two alleles are considered IBD if they are exact copies of the same allele from a single ancestor. For example, suppose the father of two siblings has genotype Aa at a particular locus. His children might both get the 'A' allele from him. If the siblings then had a child together, and they both pass down the 'A' allele, this grandchild would be AA at the locus, and their alleles would be considered IBD since they are both exact copies of the 'A' allele from the grandfather.

With this in mind, suppose we took two individuals and looked at a particular locus. At this locus, we select one allele at random from both individuals. The kinship coefficient of the two individuals at this locus is the probability that these randomly selected alleles are IBD. For example, let us consider a brother and a sister at some particular locus. We select one of the sister's alleles at random which must have come from one of her two parents. There is a 50% chance that the parent also passed that allele down to the brother, and then a 50% chance that we also pick that allele from the brother. Thus, the brother and sister have a kinship coefficient of ¼.

An individual's inbreeding coefficient, *f,* is the probability that the individual's two alleles at any locus (that is, any locus at which the individual has two alleles) are IBD. This is the kinship coefficient for the individual's parents. If the individual has only a single allele at a locus, such as with markers on the X-chromosome, nothing can be said about the inbreeding coefficient.

Null Alleles:

In general, genotyping isn't perfect and statisticians know this. There is always some probability that an individual's alleles aren't read properly, and their genotype shows up in the data as missing (we call this probability beta). This can be taken into account, and we always expect there to be some amount of missing genotypes simply due to random chance. However, sometimes these genotyping failures aren't entirely random.

One type of non-random missingness is due to problematic alleles known as null alleles. These are alleles which do not get recognized when genotyping. If an individual is heterozygous with one allele being the null allele, only the non-null allele will be read. This means that the

individual will be assumed to be homozygous for the individual's other allele. If an individual is homozygous for the null allele, their genotype will be missing. For example, suppose we are testing a sample of individuals at a certain locus, and we believe that there are two possible alleles, 'a' and 'A'. However, perhaps there is actually a third allele, 'M', which we do not recognize. Suppose Sally is MA at this locus. Sally will be read as AA. Suppose Fred is MM at this locus. Fred's genotype will be missing.

There are two primary results of the presence of null alleles. First, there will be more individuals read as homozygous than you would expect under Hardy-Weinberg Equilibrium. Secondly, there will be more missing genotypes than would otherwise occur.

This is where markers on the X-chromosome really begin to get interesting. Since males only have a single allele, if that allele is the null allele, their genotypes will ALWAYS be missing. On the other hand, females need to have two null alleles before they cause the genotype to be missing. This means that, if there is a null allele at markers on the X-chromosome, males will have far more missing genotypes than females.

Maximum Likelihood:

We have so far learned about allele frequencies, null alleles, inbreeding coefficients and, to a certain extent, beta values (random missingness). The big question now is this: how do we get good estimators when all these factors must be simultaneously considered? After all, you can't directly estimate any of these without knowing the others. You might have excess homozygosity, but how do you know whether inbreeding or a null allele is causing it? Well, it turns out we can still get decent estimators for all of these using something called a likelihood function. Furthermore, estimate all of them at the same time.

A likelihood function is a function which we will use to derive estimators with desirable properties. In our case, since our data will be discrete, the likelihood is simply the probability of the data, viewed as a function of the parameters. For example, suppose we were going to roll a die (we don't know how many sides it has) ten times and count the number of 1's we roll. Let X be the number of 1's rolled and let p be the actual probability of rolling a 1. Then $L(p|X)$, the likelihood function of p given X, takes in an estimated value of p and outputs the probability that

we would see X 1's in ten rolls, assuming that the estimated value of p is the true value of p. In the case of our example, L(p|X) would look something like this:

$$L(p|X) = \binom{10}{x} p^x (1-p)^{10-x}.$$

Suppose we get two 1's in ten rolls, and we guess at random that p is about 0.1. Then our likelihood function tells us that $L(0.1|2) = \binom{10}{2} 0.1^2 (1-0.1)^{10-2} \approx 0.19$. Thus, if p were 0.1, the chance we would get two 1's in ten rolls is roughly 0.19. One way to get a good estimator for p is to find the value, $\hat{p}$, for which the likelihood of the observed data is highest. That is, find the value $\hat{p}$ which maximizes $L(p|2)$. This is known as a maximum likelihood estimator. If you've taken any calculus, you might know that one way to do this is to take the derivative (actually the derivative of the log of the likelihood since it is easier), set it equal to zero, then solve for p. If we were to do this, we would get that $\hat{p} = 0.2$, or more generally, $\hat{p} = x/n$ where n is the number of trials and x is the number of successes. This is sometimes known as sample proportion, and it is the maximum likelihood estimator for a proportion, p.

Now, likelihood functions can have multiple parameters. In our case, we want a likelihood function for allele frequencies (including null allele frequencies), inbreeding coefficients, and beta values. We will find maximum likelihood estimators for each of these values. However, have a problem. The likelihood function we need requires us to know the IBD status of each genotype and every individual in our dataset. We don't have this information. This means we won't be able to use simple calculus. Instead, we will use something known as an EM algorithm.

EM Algorithms:

An EM Algorithm is a numerical method for finding the values of parameters which maximize a likelihood function. It allows you to get maximum likelihood estimators even when you are missing some required data. Note that, since the natural log is an increasing function over the non-negative real numbers (and all likelihoods/probabilities are non-negative real numbers), finding the values of parameters which maximize the log of the likelihood function is the same as finding the values of parameters which maximize the original likelihood function. We often try to maximize the "loglikelihood function" instead of the normal likelihood function as, perhaps counterintuitively, it is usually much simpler.

The idea behind the EM algorithm is to work with the "complete data likelihood" rather than the "observed data likelihood." For example, in a simplified situation, suppose we have the genotypes of a single individual at several markers and we are trying to get a maximum likelihood estimator for his inbreeding coefficient, *f*. The likelihood function given only the observed genotypes is impractical to work with, but it would be far simpler if we also knew the IBD status at each of his markers. This would be the complete data. The likelihood function given both the genotypes and the IBD status is called the complete data likelihood function.

The EM Algorithm works by starting with a guess for each of the parameters and repeating a two-step loop. In the simplified example, we would start with a guess for *f*. Next, we perform the E-step. We take the expected value of the loglikelihood function with respect to the complete data given the observed data and the current guess for the parameters. In the simplified example, we would take expected value of the IBD statuses at each marker given the guess for *f* and the observed genotypes. This expected value will still be a function of *f*. For the M-step, we find the parameter values which maximize the expected value of the complete data. In the simplified example, we would maximize the expected IBD statuses with respect to *f*. This gives us a new, slightly closer estimate for *f*.

If we repeat this loop thousands of times then, provided we started with a close enough guess, we should eventually converge to the maximum likelihood estimators. This whole process has already been researched and implemented, with respect to the genetics issues outlined in this paper, by Amy Anderson and her co-workers. However, there is still an issue with their work.

X-Linked Markers:

One area where the EM-algorithm has trouble is markers on the X chromosome. These were mentioned earlier in this paper. At these markers, females have two alleles where males have only a single allele. The software developed by Amy and her co-workers is unable to deal with males at X-linked markers. If you want to do any kind of testing on X-linked markers, you are forced to use a dataset of entirely females. Furthermore, there is relatively little research on the effects of using males at X-linked markers. This leaves a few interesting concepts to be explored if we can get the EM-algorithm to work properly with X-linked markers.

Recall from the section on null alleles that if an individual is heterozygous for a null allele at a marker, their genotype will come out as homozygous for the other allele, and if they are homozygous for a null allele, their genotype will be missing. Males at X-linked markers only have a single allele: there is no concept of heterozygosity or homozygosity. They only have to get the null allele once to come out missing. This means that males at X-linked markers should have far more missingness due to null alleles than you would see in females at the same markers. Furthermore, there can't be any excess homozygosity due to inbreeding coefficients or null alleles. This means that in males at X-linked markers, null alleles should stick out like a sore thumb.

This leads to some interesting ideas. Recall that the EM-algorithm attempts to estimate every parameter based on every other parameter. Improving estimation of any one parameter trickles down and improves estimation at every other parameter. Perhaps including males at X-linked markers dramatically improves the estimation of null alleles at X-linked markers, which then improves estimation of inbreeding coefficients in the females since the algorithm "knows" exactly how much excess homozygosity to attribute to null alleles, and thus how much to attribute to inbreeding. This improved estimation of inbreeding coefficients would then improve estimations at all other markers since inbreeding coefficients are specific to individuals, not markers. We believe it is possible, then, that males at X-linked markers could be a powerful tool for improving estimation of all parameters in the algorithm. For example, it might even be useful when trying to estimate inbreeding coefficients in a dataset of females to throw in a few males and include X-linked markers, even though the males have no concept of an inbreeding coefficient.

At this point, we have two main goals. First, we need to adjust the software for the EM-algorithm to function on males at X-linked markers. Secondly, we need to test the modified algorithm on thousands of simulated datasets to compare it to the old algorithm and see if males at X-linked markers affect overall results as we have hypothesized.

Modifying the EM-Algorithm:

This is where our work really begins. We modified the existing EM-algorithm to properly handle males at X-linked markers. We will not go into great detail on the coding itself, as it isn't really the point of this paper. However, we will discuss some of the main points of the process.

The algorithm itself is part of a massive, roughly 2000-line bundle of C code which contains several different methods including multiple versions of the EM-algorithm, likelihood functions, and the data simulation software. Our goal was to create new, improved versions of the old likelihood function and EM-algorithm which can handle males at X-linked markers. There were two main steps in this process. First, we modified the likelihood function to include males at X-linked markers. This was relatively simple because of the way the likelihood function is set up. We simply added a couple of extra if statements. Second, we had to modify the code for the E-step and M-step to be able to deal with X-linked males. This is where it gets tricky.

As you might imagine, C has no built-in function for taking the expected value of a function nor maximizing a function. The way the E-step and M-steps are coded involves a number of mathematical "tricks" which allow the computer to get the job done. We will not discuss these tricks here as they aren't important and only serve to overcomplicate this paper. However, we need these tricks to do what we want them to do with males at X-linked markers. This is tricky, because of the complexity of the overall program. We want to change as few things as possible. However, we found a way to deal with the X-linked males in a way that uses the existing setup of the program very nicely.

The program already knows how to deal with genotypes with only a single allele under slightly different circumstances: IBD genotypes. When an individual is IBD at some marker, the algorithm treats it as one allele for the purpose of all calculations. This is coincidentally exactly what we want to algorithm to do with males at X-linked markers. With this in mind, for the purpose of most calculations, we simply had the algorithm treat males at X-linked markers as if they were IBD. This worked in all cases except one: recalculation of the inbreeding coefficients. Adding a whole bunch of IBD genotypes would normally ruin this process. However, we conveniently set up this process to skip males at X-linked markers, since they don't directly add any information about inbreeding. In this way, everything worked out perfectly.

Simulations:

With the EM-algorithm set up to deal with males at X-linked markers, the next step was to run thousands of simulations to see how it does. As noted earlier, we want to examine how adding males at X-linked markers affects the overall estimations. With this in mind, we designed the first set of simulations in the following way:

- All simulations had 100 females, varied number of males from 0 to 50
- All simulations had 100 total markers, varied proportion of X-linked markers
- For each number of males, ran 2000 simulated datasets and averaged results across all of them
- Ran both the modified algorithm and the old algorithm for comparison, with the old algorithm ignoring X-linked markers entirely

What we found was not what we expected. The old algorithm was getting better results than ours in every case with regards to inbreeding coefficients, and worse results than ours in every case with regards to inbreeding coefficients. At first, we thought this was strange, but not unexplainable: it is possible that our algorithm is doing better on null alleles for reasons described previously, and this is actually causing an over-attribution of excess homozygosity to null alleles, thereby causing our algorithm to underestimate inbreeding coefficients. However, there was a later insight which ruins this whole idea.

Problems:

Our modified algorithm is doing worse than the old algorithm on inbreeding coefficients even when there are no males in the dataset. This is impossible. When there are no males in the dataset, X-linked markers are no different from normal markers. Our algorithm should act exactly like the old one, but with more data to work with. It couldn't possible be doing consistently worse. Because of this, we believe there is something wrong with the program. We have spent several months attempting to fix this issue, but we have had little success.

We know that the new algorithm functions exactly like the old algorithm when there are no X-linked markers. This is good, and completely expected. Furthermore, we know that the new algorithm with no males functions exactly as it does with no X-linked markers. This is also good, and completely expected. It is only when the old algorithm is run with X-linked markers in the dataset that everything seems to go wrong. However, we are out of time and must rest our research for the moment.

Conclusion:

We conclude our research with progress made, but issues remaining. We have a number of ideas as to what could be causing the bug we are observing. These will be tested at a later

date, possibly in another project. For now, we hope to leave you with a basic understanding of how the EM-algorithm works, the changes we sought to make, and the obstacles we faced.